

Note:

- used Stata and Julia
- relatively short

INVESTIGATING THE ODDS OF LOAN ACCEPTANCE BEFORE AND AFTER THE GFC

University of Auckland

Abstract

This paper investigates the factors influencing loan application success for 458,119 individuals in the US before / during the GFC and post-GFC. In addition, it also investigates the factors on which individuals are grouped and how it affects their loan application outcome. We begin by running OLS and Logistic regressions on the two time periods individually. More variables are significant in determining loan application outcome post-GFC compared to pre-GFC. A Conditional Logistic regression is implemented next to investigate factors affecting outcome of groups of individuals and the likelihood of individuals being in them. Overall, employment years and debt-to-income ratio are the two key factors in determining loan application success.

Contents

1	Introduction	1
2	Literature Review	1
3	Data and econometric model	2
3.1	Data	2
3.2	Econometric Model	3
	OLS model	3
	Logistic model	4
	Conditional Logistic model	5
4	Results and Interpretation	7
4.1	OLS	7
4.2	Logistic models	8
4.3	Conditional Logistic model	9
5	Conclusion	9
5.1	Summary	9
5.2	Limitations	10
5.3	Further Research	10
	References	11
	Appendix	12

1 Introduction

It has been nearly eight years since the GFC of 2007 - 2009, sent the worldwide economy into a turmoil; the effects of which are still being experienced in some parts of the world. Many theories have been postulated on the cause of this crisis with a key one being the collapse of the housing market [1]. This was due to sub-prime loans given out by banks, which begs the question: How easy was it to get a loan and what criteria was used to grant loans in the US? Lending Club [2] is a peer to peer lending company offering loans with low interest rates in the US. There have been claims that these interest rates for the majority have been low throughout and post-GFC.

This paper looks at the probability of a loan being granted and what factors play a role in determining the outcome of a loan application in the US. It will also investigate how these factors have changed ever since the GFC.

The analysis begins with a literature review and the contributions of this paper. The next section discusses the data source and the econometric models used. Following this, empirical results from the regressions are presented with an intuitive discussion behind these results. The final section briefly concludes and discusses some limitations of the data and methodology with a direction for future research.

2 Literature Review

Literature discussing the econometrics behind ‘loan data’ has been relatively sparse, due to confidentiality of data provided by banks and the incompleteness of individual / firm specific characteristics. Two main contributions to this area of research have been presented by Abildgren, Drejer and Kuchler (2012) [3] and Munnell, Tootell, Browne and McEneaney (1996) [4].

The former article is an analysis of banks’ loan rejection rates and the creditworthiness of the bank’s corporate customers. A probabilistic model is constructed using data on solvency and profit ratios and some firm-specific characteristics for firms in Denmark. The model is esti-

mated separately for the years 2007 and 2009 / 2010. Denmark presents an interesting case to investigate loan rejection rates because even though it had recessionary effects due to the GFC, its effects were lower than its neighbouring countries due to its strong fiscal position and sound fiscal policy framework. Consequently, the results from the paper show that even during the financial crisis 2009 / 2010, only firms with weak economic performance had low bank loan acceptance rates.

The latter article discusses mortgage lending in Boston. In particular they look at the probability of being granted a loan for the different races residing in Boston in 1990. An OLS and a logistic model were constructed factoring in individual characteristics such as gender, race, age, marital status and number of dependants with a couple of behavioural results. The first one being that black and Hispanic mortgage applicants in the Boston area were more likely to be turned down than white applicants with similar characteristics. However, a more interesting result being that none of the lenders indicated using race as a signal for loan determination. Hence, there are some fixed unobserved effects not captured by these techniques. This investigation looks to account for these effects.

3 Data and econometric model

3.1 Data

The data consists of characteristics of 458,119 individuals in the United States who applied for a loan during 2007 to 2009 i.e. during the GFC and in 2011 i.e. post-GFC. In the strict sense, 2011 is not entirely after the GFC since the effects were still being experienced worldwide. However, by 2011, the US had started to implement tighter credit policies to save their economy hence, it presents an interesting case to investigate lender reactions during this period. The composition of the sample was determined by data availability to ensure as complete a sample was available.

All data was obtained from the Lending Club database. Each individual who applied for a loan has a binary classification indicating whether the loan application was accepted (1) or rejected

(0). For majority of the individuals who applied for a loan either the loan amount was granted in full or not at all. Thus, the sample only focuses on these individuals. The data for individuals who were denied a loan did not consist of their income levels however, the debt-to-income ratio for all individuals was given and this has been used as a substitute. The number of years of employment for each individual is an aggregate amount of years an individual has been employed regardless of any periods of unemployment. Each individual's State of residence when they applied for a loan has also been provided.

3.2 Econometric Model

OLS model

Following a similar technique to Munnell et. al (1996), two OLS models are initially created for individuals who applied for a loan during GFC and for those who applied for a loan post-GFC. The specification is as follows:

$$Y_i = \beta_0 + \beta_1 loan_i + \beta_2 DTI_i + \beta_3 emp_i + \beta_4 loanDTI_i + \beta_5 loanemp_i + \beta_6 DTIemp_i + \beta_7 loanempDTI_i + \epsilon_i \quad i = 1, 2, \dots, N$$

where

$$Y_i = \begin{cases} 1, & \text{individual } i\text{'s loan granted} \\ 0, & \text{individual } i\text{'s loan rejected} \end{cases}$$

$loan_i$ is individual i 's loan request, DTI_i is individual i 's debt-to-income ratio and emp_i is individual i 's years of employment. Since variables such as annual income are missing, interaction terms with coefficients β_4 , β_5 , β_6 and β_7 have been created to minimise the omitted variable bias.

Logistic model

As I will discuss later, the OLS models above give results which are simple to interpret. However, since the dependent variable is binary, we're interested in estimating the probability or odds of a loan being accepted where the odds lie between 0 and 1. This leaves us with either a choice of a *probabilistic* or a *logistic* model [5]. Both models give relatively consistent results however the distribution of the regressors in a logistic model have thinner tails compared to a probabilistic model. In our scenario, majority of the loan applicants appear to be middle-class working individuals. This implies for example, the distribution of loan amount requested or income will contain majority of individuals in a similar range with a few outliers, i.e. few individuals with small / large loan requests or low / high incomes, which means the distribution will contain a thin tail [5]. For this reason, a *logistic* model has been implemented. Using the same regressors as specified in the OLS model, the specification assumes there is a latent variable Y^* such that:

$$Y^* = \mathbf{X}\beta + \epsilon$$

where

$$y_i = \begin{cases} 1, & \text{if } y_i^* > 0 \quad i = 1, 2, \dots, N \\ 0, & \text{if } y_i^* \leq 0 \quad i = 1, 2, \dots, N \end{cases}$$

The variable Y^* is latent because we do not observe ϵ . This is also estimated individually for the two time periods. This logistic model follows a logistic distribution with the logistic function defined as

$$F(x) = \frac{1}{1 + e^{-\mathbf{X}\beta}}$$

The coefficients of the logistic models were estimated using two maximum likelihood techniques: the Iterative Reweighted Least Squares (IRLS) and Generalised Linear Models (GLM) algorithms. Following this, the marginal effects from the logistic models were also determined.

Conditional Logistic model

As mentioned earlier, previous research goes as far as either estimating an OLS, probabilistic or logistic model for either corporate loans or personal loans for individual time periods. Estimating probability or odds of a loan being granted over time becomes difficult since the individuals who applied for a loan in one period may not have applied for a loan in another time period (this is often the case for majority of individuals). In addition, tracking data for each individual's loan applications over time can be a challenge in itself. Even if this data could be tracked, there could be an individual characteristic unobserved over time.

In light of these concerns, the data set used in this exercise also does not contain the same individuals over the two time periods. So, rather than focusing on the odds of loan acceptance and the factors that influence it over time, we can focus on the odds of loan acceptance and the factors that influence it across different groups or *clusters* of individuals who share some same unobserved characteristic or a *fixed effect* within each cluster [5][6]. In our scenario examples of a fixed effect within clusters could be job experience, intelligence and genetic makeup.

A total of 4000 clusters were created containing individuals from both time periods. To create the clusters loan applicants during the GFC were considered first. Using a technique known as *K-means clustering*[7], these applicants were separated into 4000 clusters with an optimum separation criteria decided by the software 'Julia' such that there is an even spread of individuals across the clusters. After the clustering is done, the cluster number in which each individual is assigned, the number of assignments in each cluster and the centroid of each cluster is generated. An example of the cluster representation of our data with 5 clusters is shown in Figure 1.

To maintain consistency and accuracy, the same 4000 centroids were then used to create clusters for loan applicants post-GFC. An example of the cluster representation of our data post-GFC using the same 5 centroids as the above example is shown in Figure 2.

Now that each individual from both time periods has been classified into one of 4000 clusters, we now investigate the odds of loan acceptance and factors affecting loan acceptance across individuals within clusters, accounting for any fixed effects within each cluster. This is done through a *conditional logistic model*. The model allows us to investigate how the characteristics

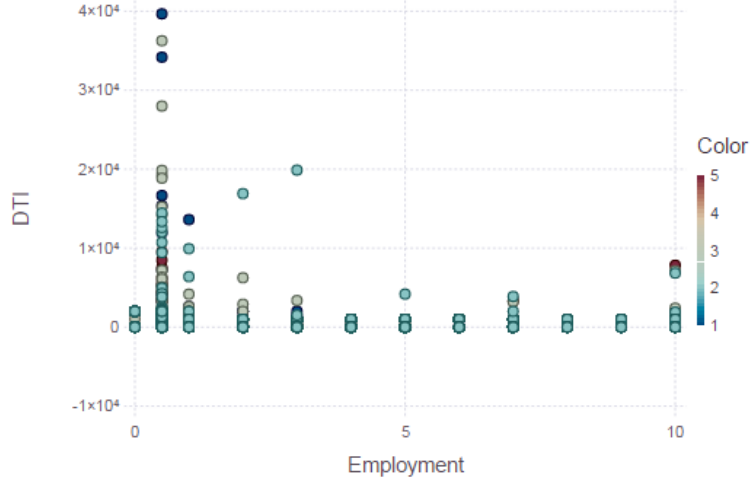


Figure 1: GFC loan applicants split into 5 clusters

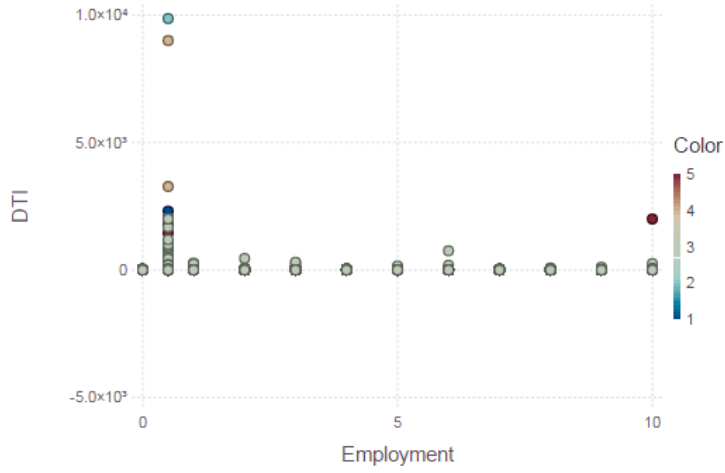


Figure 2: Post-GFC loan applicants split into 5 clusters

of the clusters affect individuals likelihood of being in them. It also allows us to control for any unobserved characteristics of the clusters which remain fixed for all individuals within the clusters. This is analogous to a panel regression except the groups are clusters and we look at the effects across individuals in the clusters rather than time. The model can be specified as follows:

$$\begin{aligned}
 Y_{ci} = & \alpha_0 + \alpha_1 loan_{ci} + \alpha_2 DTI_{ci} + \alpha_3 emp_{ci} + \alpha_4 loanDTI_{ci} + \alpha_5 loanemp_{ci} \\
 & + \alpha_6 DTIemp_{ci} + \alpha_7 loanempDTI_{ci} + \epsilon_{ci} \quad i = 1, 2, \dots, N \quad c = 1, 2, \dots, C
 \end{aligned}$$

where

$$Y_{ci} = \begin{cases} 1, & \text{individual } i \text{ in cluster } c \text{'s loan granted} \\ 0, & \text{individual } i \text{ in cluster } c \text{'s loan rejected} \end{cases}$$

In this scenario, the panel is unbalanced. The regression was estimated using the GLM algorithm. Following the estimation of this model, the odds ratios were also calculated.

4 Results and Interpretation

4.1 OLS

Starting with loan applicants before / during the GFC, we find that the number of years of employment has a strong significance whereas, interaction terms between the loan amount, number of years of employment and debt-to-income ratio have weak / some significance in determining the probability of loan acceptance. The significance of the interaction terms suggests there were omitted variables in the original data set which are being accounted for with these interaction terms. With a 1 year increase in employment years, the probability of loan acceptance on average increases by 1.8 percentage points, whereas a unit increase in the interaction terms has a negligible change on average on the probability of loan acceptance. Results are summarised in Table 1 in the Appendix.

For loan applicants post-GFC, we find that the loan amount requested, debt-to-income ratio, employment years and all their interaction terms have a strong significance in determining the probability of loan acceptance. Again, the significance of the interaction terms suggests there were omitted variables in the original data set which are being accounted for with these interaction terms. With a 1 year increase in employment, the probability of loan acceptance on average increases by 2.5 percentage points. A 0.1 increase in the debt-to-income ratio reduces the probability of loan acceptance on average by 0.024 percentage points. A more interesting result is that a one dollar increase in the loan amount requested negligibly (virtually no change) reduces the probability of loan acceptance on average, even though this term is deemed signif-

icant. Results are summarised in Table 2 in the Appendix. Overall, these results are consistent with our theory that post-GFC, the tightening of credit policies has put more weight on all the variables in determining loan acceptance as opposed to before / during the GFC.

While these results seem intuitive, OLS method provides *percentage point* changes, which means the probability can go below 0 or above 1. Due to this reason, the results from the logistic models are more reliable.

4.2 Logistic models

For loan applicants before / during the GFC, both the IRLS and GLM algorithms gave the same results. The coefficients from the logistic regression suggest that employment years and the interaction term between loan amount requested and employment years have a strong significance in determining odds of loan acceptance. The sign of the coefficient on these terms only is useful in determining the direction of causality. The marginal effects of the coefficients are more useful. A one year increase in employment years increases the odds of loan acceptance on average by 1.2% whereas, a one unit increase in the loan amount-employment years interaction term reduces the odds of loan acceptance on average negligibly. This result is consistent with the OLS result where employment years seemed to be the primary factor in determining loan acceptance. The results are summarised in Table 3 in the Appendix.

For loan applicants post-GFC, the IRLS and GLM algorithms produced different results. The IRLS algorithm successfully converged but the GLM algorithm failed to converge. Nonetheless, results from both algorithms are provided in Table 4 and Table 5 in the Appendix. For statistical inference we'll use results from the IRLS algorithm since it successfully converged. The loan amount requested, debt-to-income ratio, employment years and all their interaction terms have a strong significance in determining the odds of loan acceptance. A 0.1 increase in the debt-to-income ratio reduces the odds of loan acceptance on average by 0.033%. With a dollar increase in the loan amount requested, the odds of loan acceptance on average reduce negligibly. However, a more interesting result is a 1 year increase in employment *reduces* the odds of loan acceptance on average by 0.06% (the effect is opposite with the GLM results although

the GLM results are not reliable). This seems counter intuitive as the odds should increase. But with the reduction being so low, it may just be a property of the data set or we can interpret this as a negligible effect.

Overall these results are consistent with the OLS results and consistent with logic / theory. The tightening of credit policies has put more weight on all variables when determining the outcome of a loan application post-GFC compared to before / during the GFC.

4.3 Conditional Logistic model

After individuals from both time periods were classified into clusters, the conditional logistic model was implemented. The results indicate that within each cluster, debt-to-income ratio and employment years have a strong significance in determining loan acceptance across individuals while none of the interaction terms are significant. Again, the sign of the coefficients is only used for direction of causality and we look to the odds ratios for inference. A 1 year increase in employment *multiplies* the odds of loan acceptance on average by 1.2 or increases the odds by 20%. With a 0.1 increase in the debt-to-income ratio, the odds of loan acceptance decrease on average by 1%. The results are summarised in Table 6 in the Appendix

Given the characteristics of the clusters produced, the individual's employment years and debt-to-income ratio affect the likelihood of being allocated to a cluster.

5 Conclusion

5.1 Summary

The different techniques applied in this investigation roughly give the same intuitive results. More variables help determine loan acceptance post-GFC compared to before / during GFC. As mentioned earlier, this intuitive result is due to tighter credit policies employed by lenders. In both time periods, the number of years of employment and debt-to-income ratio stand out as the two key factor in determining loan application success. The logic behind this could be that the

more number of years an individual is employed, the greater amount of income they will have saved to make repayments. Similarly, the lower the debt-to-income ratio, the lower the chances of an individual defaulting on their loan repayments. A final interesting result is that groups of individuals seeking loans from Lending Club have had their outcome judged primarily on their employment years and debt-to-income ratios after accounting for any latent fixed effects. Whether Lending Club did this on purpose or subconsciously is another question.

5.2 Limitations

- There may be a potential omitted variable bias since we were restricted to a few variables with interaction terms used to minimise this bias.
- Employment categories were only classified up to 10 years with the remainder as 10+ years. Additional categories could have improved inference.
- Conditional logistic results for post-GFC data are not as reliable due to different methods producing different results.
- Available computation power. With the available computation power, the total number of clusters was restricted to 4000. More powerful computers could compute more clusters which could again improve our inference.

5.3 Further Research

- Different clustering techniques and different clusters should be implemented and compared to make inference more robust.
- Other maximum likelihood techniques should also be considered and compared after clustering.
- As we found, logistic regression techniques can sometimes give unreliable results. Additional econometric techniques should be investigated.
- Probabilistic regressions should also be performed and compared with the logistic regressions.

References

- [1] Justine Davies. *Global Financial Crisis – What caused it and how the world responded*.
<http://www.canstar.com.au/home-loans/global-financial-crisis/>
- [2] Lending Club. <https://www.lendingclub.com/info/download-data.action>
- [3] Kim Abildgren, Peter A. Drejer, Andreas Kuchler. *A micro-econometric analysis of the banks' loan rejection rates and the creditworthiness of banks' corporate customers*. Working Paper. Danmarks Nationalbank.
- [4] Alicia H. Munnell, Geoffrey M.B. Tootell, Lynn E. Browne, James McEneaney. Mortgage Lending in Boston: Interpreting HMDA Data. *The American Economic Review*, 86(1):25-53, 1996.
- [5] Cameron, A., & Trivedi, P. (2005). *Microeconometrics Methods and Applications*. New York, USA: Cambridge University Press.
- [6] German Rodriguez. *Generalized Linear Models*.
<http://data.princeton.edu/wws509/notes/c6s3.html>
- [7] *K-means Clustering*.
<http://www.onmyphd.com/?p=k-means.clustering>

Appendix

Variable	Coefficient	s.e.	T-statistic
Loan	-6.11e-8	9.24e-8	-0.66
DTI	-1.82e-6	4.52e-6	-0.40
Emp	0.018	0.00031	56.55
Loan*DTI	9.14e-11	3.06e-10	0.30
Loan*Emp	-2.83e-7	2.20e-8	-12.89
Emp*DTI	-3.73e-6	1.99e-6	-1.87
Loan*Emp*DTI	-2.57e-10	1.29e-10	-1.99

Table 1: OLS before / during GFC

Variable	Coefficient	s.e.	T-statistic
Loan	-9.52e-7	5.48e-8	-17.39
DTI	-0.0024	1.74e-5	-137.40
Emp	0.025	0.00034	73.52
Loan*DTI	8.94e-8	9.37e-10	95.39
Loan*Emp	5.75e-7	1.87e-8	30.67
Emp*DTI	0.00048	3.21e-5	149.53
Loan*Emp*DTI	-1.79e-7	1.72e-9	-104.06

Table 2: OLS post-GFC

Variable	Coefficient	s.e.	Z-statistic	MargEffect(acc)	MargEffect(rej)
Loan	-5.89e-6	1.51e-6	-3.90	-4.04e-7	4.04e-7
DTI	-0.00012	0.00014	-0.83	-8.06e-6	8.06e-6
Emp	0.17	0.0032	54.32	0.012	-0.012
Loan*DTI	-1.25e-8	1.25e-8	-1.00	-8.58e-10	8.58e-10
Loan*Emp	-1.62e-6	2.44e-7	-6.62	-1.11e-7	1.11e-7
Emp*DTI	-1.07e-5	3.00e-5	-0.36	-7.35e-7	7.35e-7
Loan*Emp*DTI	-3.02e-9	2.38e-9	-1.27	-2.07e-10	2.07e-10

Table 3: Logistic before / during GFC

Variable	Coefficient	s.e.	Z-statistic	MargEffect(acc)	MargeEffect(rej)
Loan	-3.04e-5	1.30e-6	-23.47	-2.19e-6	2.19e-6
DTI	-0.047	0.00016	-288.68	-0.0034	0.0034
Emp	-0.0085	0.0038	-2.24	-0.00061	0.00061
Loan*DTI	-2.58e-6	1.02e-8	-253.98	-1.86e-7	1.86e-7
Loan*Emp	2.19e-6	2.32e-7	9.42	-1.57e-7	1.57e-7
Emp*DTI	0.094	0.00018	517.02	0.0068	-0.0068
Loan*Emp*DTI	5.16e-6	9.58e-9	538.86	3.71e-7	-3.71e-7

Table 4: Logistic post-GFC using IRLS algorithm

Variable	Coefficient	s.e.	Z-statistic	MargEffect(acc)	MargeEffect(rej)
Loan	-0.000011	1.10e-6	-9.57	-7.70e-7	7.70e-7
DTI	-0.040	0.00051	-78.80	-0.0029	0.0029
Emp	0.22	0.0039	56.42	0.016	-0.016
Loan*DTI	1.05e-6	2.14e-8	48.98	7.70e-8	-7.70e-7
Loan*Emp	3.91e-6	2.19e-7	17.82	2.86e-7	-2.86e-7
Emp*DTI	0.079	0.00098	81.33	0.0058	-0.0058
Loan*Emp*DTI	-2.10e-6	4.07e-8	-51.64	1.54e-7	-1.54e-7

Table 5: Logistic post-GFC using GLM algorithm

Variable	Coefficient	p-value	Odds Ratio
Loan	-8.66e-6	0.380	0.99
DTI	-0.0061	0.000	0.99
Emp	0.182	0.000	1.20
Loan*DTI	1.36e-8	0.873	1.00
Loan*Emp	1.44e-6	0.201	1.00
Emp*DTI	-0.00011	0.627	0.99
Loan*Emp*DTI	2.77e-9	0.833	1.00

Table 6: Conditional logistic